What Causes Variation in Country-Level Life Expectancies?

STAT 462 Final Project Report

Claire Kessel

August 10, 2023

Abstract

This analysis uses health factor data from the World Health Organization (WHO) in conjunction with economic data from the United Nations website. The complete dataset, which was sourced from Kaggle, contains 22 variables about 183 countries from years 2000 to 2015. The primary goal of this project was to construct a model that accurately explains a country's life expectancy and use it to identify key factors in life expectancy improvement. A secondary goal was to determine if either BMI or childhood health factors are significantly related to life expectancy. Through model selection, the key predictors of a country's life expectancy were found to be deaths per 1,000 people aged 15 to 60 years, health expenditure as a percentage of total government spending, income composition of resources, and HIV/AIDS-related deaths among 0 to 4-year-olds per 1,000 births. BMI was significantly related life expectancy, but childhood health factors were not.

I. Introduction

In this paper, methods and techniques learned in STAT 462: Applied Regression Analysis are utilized. A dataset containing variables related to life expectancy will be used, which was retrieved from Kaggle (Kumarrajarshi, 2018). It contains 20 predictor variables related to health and economic status and a life expectancy response variable for 183 countries. The data was collected from 2000 to 2015 by the Global Health Observatory (GHO) data repository under the World Health Organization (WHO), and its corresponding economic data was collected by the United Nations. The predictor variables include the following:

- country (nominal): country
- year (categorical): year (2000-2015)
- status (binary): developed or developing status
- adult_mortality (quantitative): deaths per 1,000 people aged 15 to 60 years
- infant_deaths (quantitative): deaths per 1,000 infants
- alcohol (quantitative): alcohol consumption per capita (ages 15+) in liters of pure alcohol
- percentage_expenditure (quantitative): health expenditure as a percentage of gross domestic product per capita
- hepatitis_B (quantitative): percentage of hepatitis B immunization among 1-year-olds
- measles (quantitative): number of reported measles cases per 1000 people
- BMI (quantitative): average body mass index
- under_5_deaths (quantitative): deaths per 1,000 children aged 0-5
- polio (quantitative): percentage of polio immunization among 1-year-olds
- total_expenditure (quantitative): health expenditure as a percentage of total government spending
- diptheria (quantitative): percentage of diptheria tetanus toxoid and pertussis immunization among 1-year-olds
- HIV_AIDS (quantitative): HIV/AIDS-related deaths among 0 to 4-year-olds per 1,000 births
- GDP (quantitative): gross domestic product per capita in USD
- population (quantitative): population of the country

- thinness_1_19 (quantitative): prevalence of thinness in people ages 1 to 19 as a percentage
- income_comp (quantitative): human development index in terms of income composition of resources ranging from 0 to 1
- schooling (quantitative): number of years of schooling

The primary goal is to determine which health and economic factors are most useful in explaining life expectancy. To achieve this, a linear regression model will be constructed and utilized to make conclusions about the health and economic variables. The questions I hope to answer in this analysis include: 1. Which health and economic factors are important in explaining life expectancy? 2. After accounting for other predictors, does a significant linear relationship exist between BMI and life expectancy? 3. After accounting for other predictors, does a significant linear relationship exist between childhood health and life expectancy?

II. Exploratory Analysis

An initial exploration of the data was made to assess the overall behavior of variables, identify any patterns, and spot potential issues. Some missing data values were found upon inspection, appearing mostly in the year 2015. Using the most recent year in this analysis initially seemed ideal since the intended strategy was to focus on a single year, but 2014 was used instead due to its relative completeness. No evident errors were found in the dataset, and remaining NA values were omitted using na.omit(). Missing values were common in countries like Vanuatu, Tonga, Togo, Cabo Verde, etc., and occured mostly in population, hepatitis_B, and GDP variables. After subsetting the data and handling missing values, 22 variables and 131 countries remain.

The descriptive statistics are provided below. Looking at the five number summary of the response variable, the mean life expectancy is around 71 years. The minimum and maximum life expectancies are about 48 and 89 years, respectively. A frequency table was made for status, providing a visualization of its binary nature. There are 19 developed countries and 112 developing countries in the dataset.

Table 1: Summary statistics of country-level life expectancies

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
48.1	64.65	72	70.51985	75.8	89

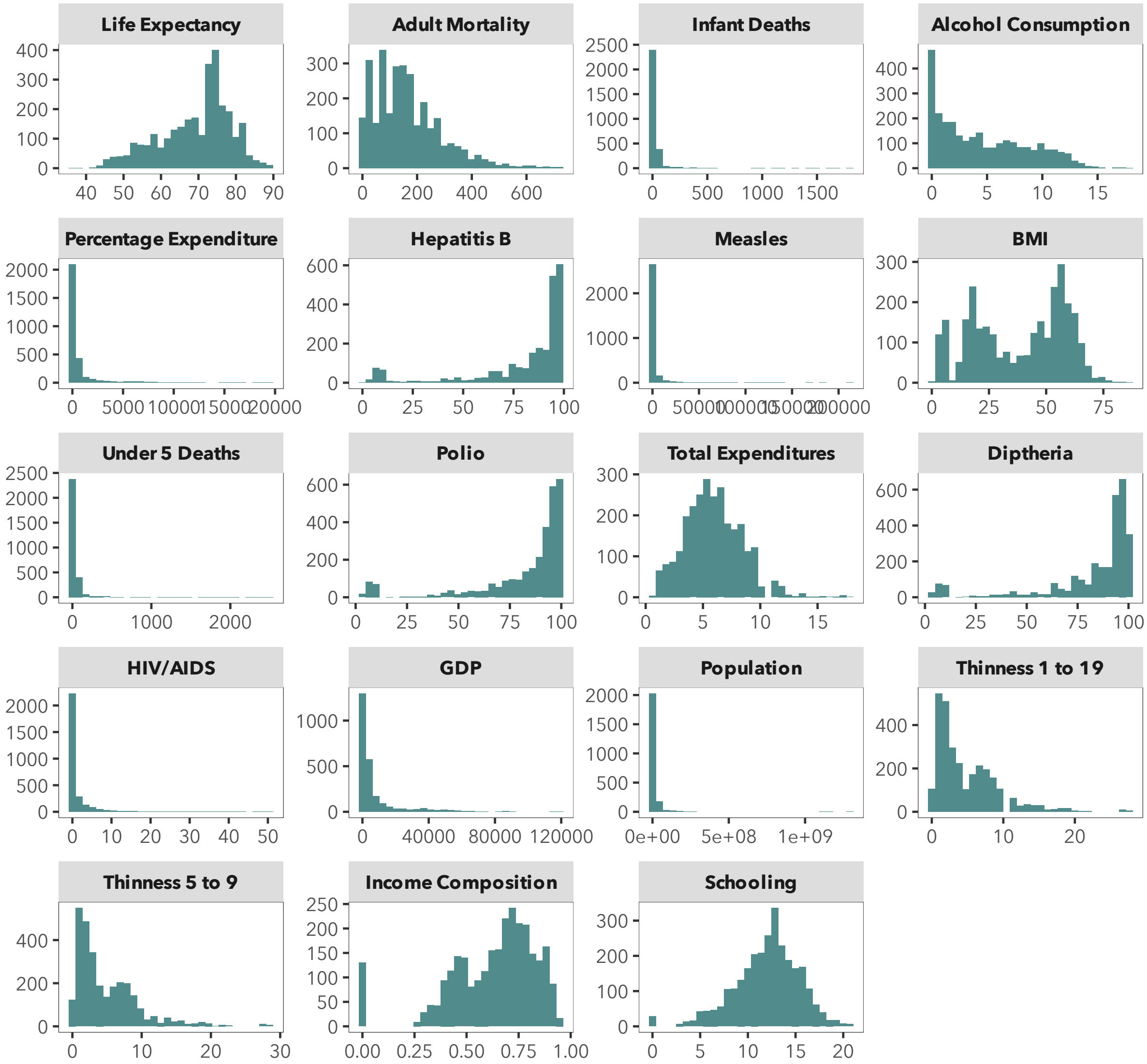
Table 2: Frequency table of development status

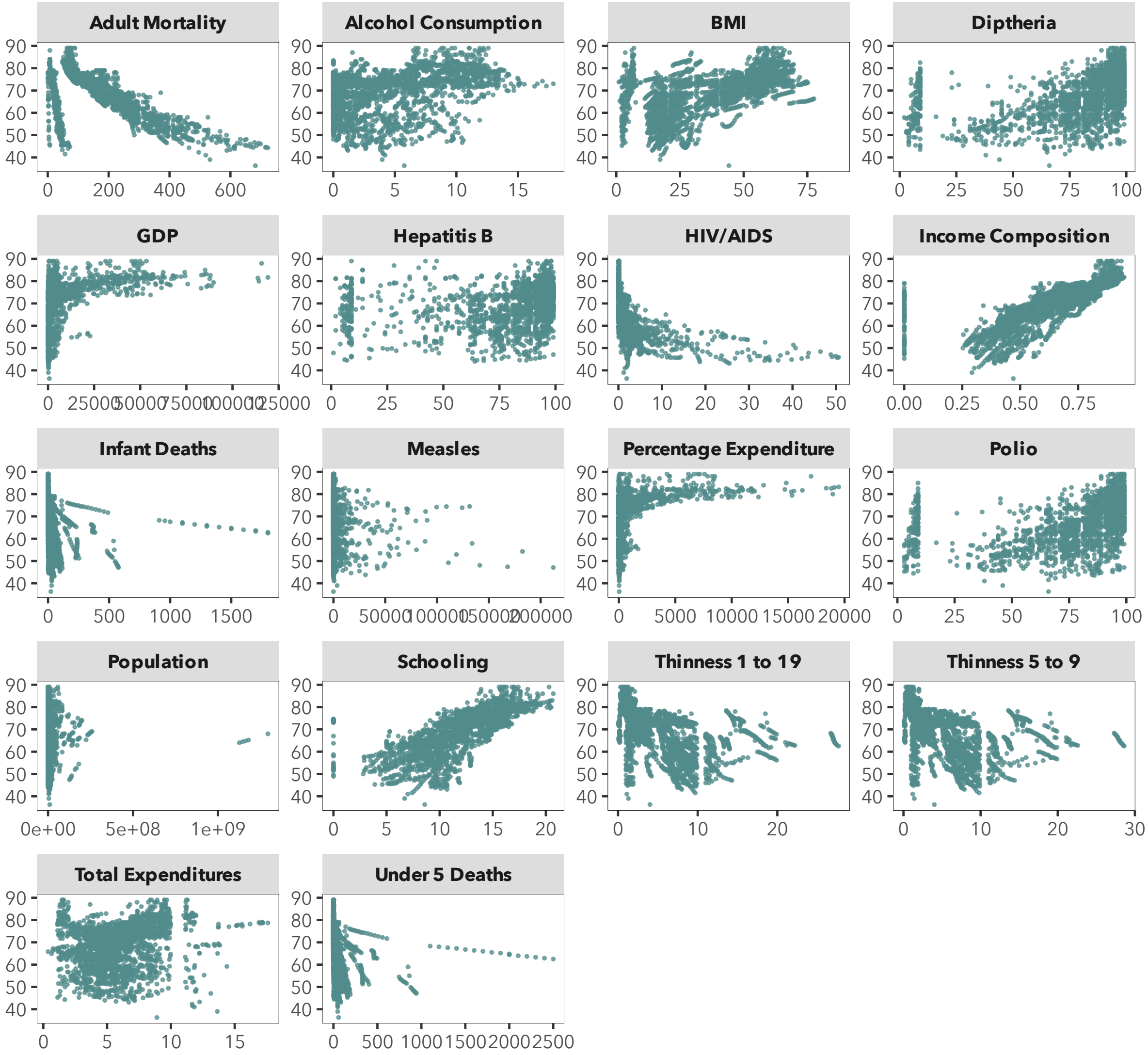
Developed	Developing
19	112

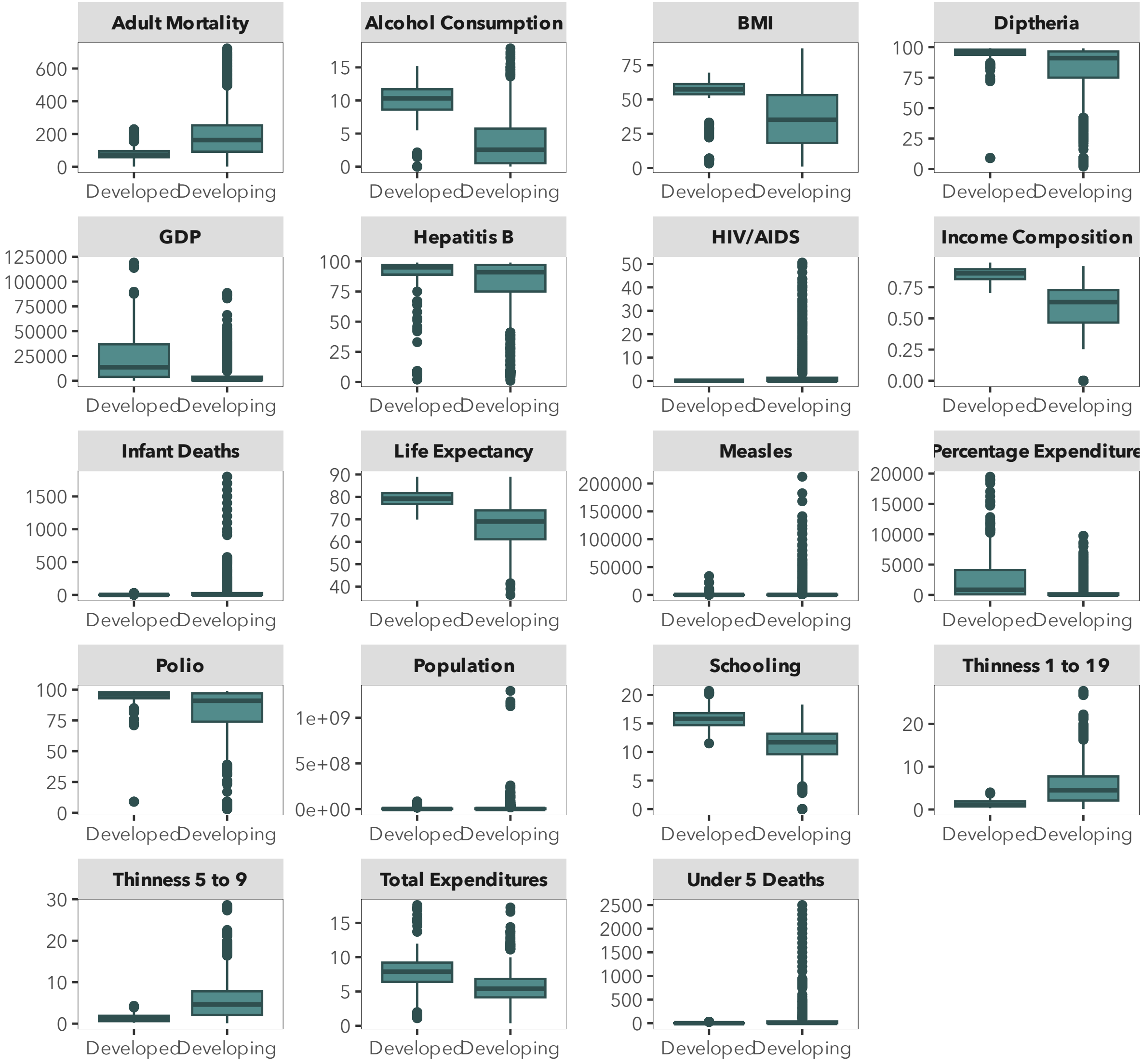
Histograms were made for each quantitative predictor. Variables related to mortality, alcohol consumption, disease cases, thinness, and healthcare expenditure were all right skewed, as expected. The histogram for GDP is also skewed right, as expected. This may be due to the far greater number of developing countries in the world versus developed countries. Considering the developing country majority in the dataset, this variable was expected to display more left skew. Total expenditures, BMI, and schooling appear to vary between countries more than other variables. It is expected for more countries to have high vaccination rates than otherwise, so the left skew in variables like diptheria, polio, and Hepatitis B makes sense.

Scatterplots of each quantitative variable against the response variable were created for pattern identification, later confirmed by a correlation matrix. adult_mortality, income_comp and schooling all show very clear linear relationships with life expectancy. BMI and total_expenditure appear to have moderately strong linear relationships with life expectancy as well. The correlation coefficients for adult_mortality, income_comp, and schooling against life expectancy were -0.77, 0.89, and 0.80, respectively. These are moderately strong. The correlation coefficient for BMI and life expectancy is 0.56, and the coefficient for total_expenditures and life expectancy is 0.32. The correlation coefficient of -0.62 for HIV_AIDS against life expectancy is surprising since a linear relationship is not extremely evident in the scatterplot. Boxplots relating the one categorical variable, status, to all quantitative variables were created in addition to these scatterplots to explore more potential relationships. These show the side-by-side comparison of quantities for developed vs developing countries.

It can be concluded from this initial inspection that that not all variables appear to be directly related to the response. In the following section, an attempt at model selection will be made.




Adult Mortality
Alcohol Consumption
BMI
Diptheria
GDP
Hepatitis B
HIV/AIDS
Income Composition
Infant Deaths
Measles
Percentage Expenditure
Polio
Population
Schooling
Thinness 1 to 19
Thinness 5 to 9
Total Expenditures
Under 5 Deaths
90
80
70
60
50
40
0 200 400 600
0 5 10 15
0 25 50 75
0 25 50 75 100
0 25000 50000 75000 100000 125000
0 25 50 75 100
0 10 20 30 40 50
0.00 0.25 0.50 0.75
0 500 1000 1500
0 50000 100000 150000 200000
0 5000 10000 15000 20000
0 25 50 75 100
0e+00 5e+08 1e+09
0 5 10 15 20
0 10 20
0 10 20 30
0 5 10 15
0 500 1000 1500 2000 2500



III. Methodology

To address the goal of understanding which health and economic factors are most useful in explaining life expectancy, a linear regression model of life expectancy was fit. The response, life expectancy, was regressed initially on all 19 predictors. In the summary output below, significant p-values at the 5 percent level are present for adult_mortality, total_expenditure, HIV_AIDS, and income_comp variables. This suggests a significant linear relationship between the predictors and the response variable.

In Figure 1, a residuals versus fitted values plot for the full model can be seen, where a random scattering of points above and below the line suggests a constant error variance. These points are centered around the horizonal line at 0, indicating a linear pattern in the model.

Figures 2 and 3 show a Normal QQ plot and histogram of the residuals, where a bit of skew is apparent at each tail. Only a handful of points deviate from the normal Q-Q line, which may be indicative of outlying or high-leverage points. The residuals seem normal nonetheless, since most points fall directly on the normal Q-Q line.

To determine whether childhood health variables like infant_deaths, under_5_deaths, thinness_1_19, and thinness_5_9 play a role in explaining life expectancy, the full model assumptions must be satisfied. A VIF table was made initially, where many variables show high multicolinearity (vif > 4). Variables with high multicolinearity and low correlation with life expectancy were removed until each vif score was below 4. In Figure 4, the residuals vs fitted values plot for this reduced model is shown. A random scattering above and below the horizontal line at zero is evident, suggesting constant error variance. These points are centered around zero as well, indicating a linear trend in the model. The normal Q-Q plot can be seen in Figure 5 along with a histogram of the residuals in Figure 6. There is some slight left skew in the residuals, but most of the points follow the line closely. The histogram shows an approximately normal distribution of residuals. Now that the model has satisfied all assumptions, it can be used to see whether childhood health factors and BMI are significantly related to life expectancy. It is evident that childhood health does not play a significant role in explaining life expectancy since neither thinness_5_9 nor deaths_under_5 variables have significant p-values in this model. BMI, however, did have a significant p-value corresponding to its individual t-test. This indicates a significant linear relationship between BMI and life expectancy. After making inferences about the relationships between BMI and childhood health factors with life expectancy, the search for the best model to explain life expectancy can proceed.

Using regsubsets(), the best model containing our 20 variables of interest was identified. An adjusted r-squared added variable plot, seen in figure 7, was then used to determine which combination of predictors produced the highest adjusted r-squared. The resulting model contained adult_mortality, total_expenditure, HIV_AIDS, and income_comp variables, which are the same four significant predictors in the full model. Yielding an adjusted r-squared value of 0.8701, this model explains 0.72 percent more of the variability in life expectancy than the initial model.

However, upon fitting this model, two potential outliers were observed. The 12th and 108th observations seen above the blue line in Figure 8 have studentized residuals of 3.32 and -3.68, respectively, indicating that these countries have both higher and lower life expectancies than expected. The 108th observation is a high leverage point as well, seen above the line in figure 9, along with the 125th, 39th, 114th, and 68th obervations. These observations have leverage values of 0.14, 0.16, 0.17, 0.17, and 0.29, respectively, where the typical cutoff for high leverage is 3p/n = 0.11.

To determine if these observations influence fit, we use the difference in fits method. 7 influential points were found, which can be seen above the blue line in figure 10. These are the 99th, 4th, 26th, 68th, 12th, 39th, and 108th observations. These were removed with caution that some lurking variable may be accounting for these differences.

Figure 1: Residuals vs Fitted Values

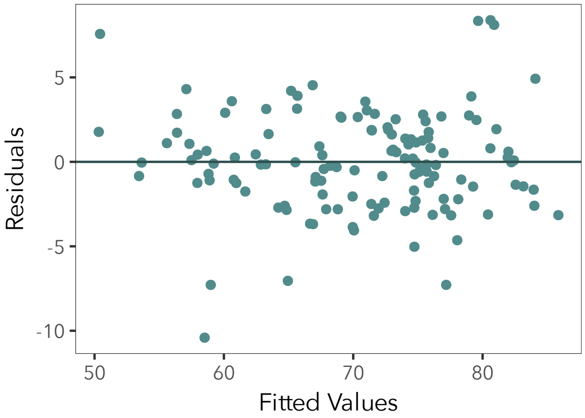


Figure 2: Normal Q-Q Plot

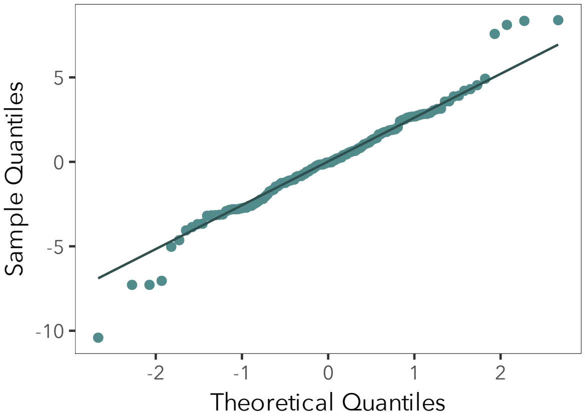


Figure 3: Histogram of Residuals

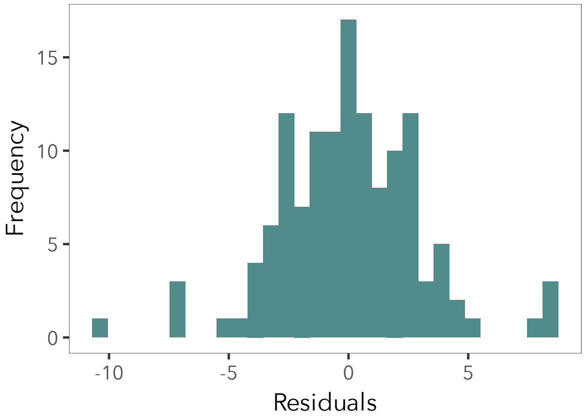


Table 3: Variance Inflation Factors

Variable	Model 1	Model 2
Adult Mortality	2.67	2.02
Alcohol Consumption	2.04	1.56
BMI	2.20	1.74
Developing Status	1.71	1.48
Diptheria	7.25	2.21
GDP	12.33	
Hepatitis B	5.70	
HIV/AIDS	1.91	1.85
Income Composition	11.37	
Infant Deaths	405.00	
Measles	2.88	1.90
Percentage Expenditure	11.82	
Polio	2.52	2.45
Population	8.07	
Schooling	7.27	
Thinness 1 to 19	12.86	
Thinness 5 to 9	13.11	2.28
Total Expenditure	1.33	1.18
Under 5 Deaths	325.13	2.35

Figure 4: Residuals vs Fitted Values

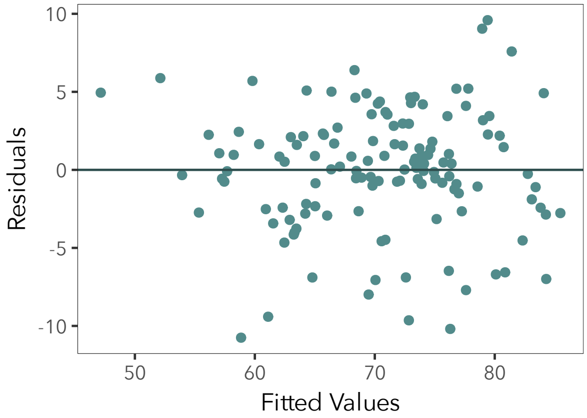


Figure 5: Normal Q-Q Plot

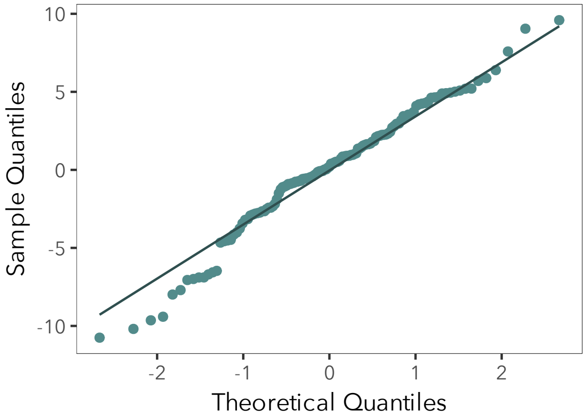


Figure 6: Histogram of Residuals

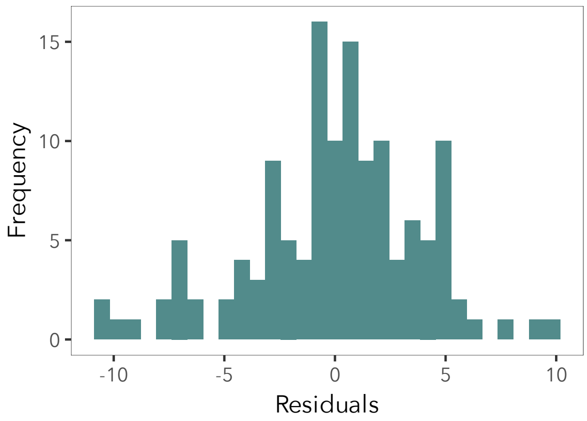


Figure 7: Adjusted R-Squared Added Variable Plot

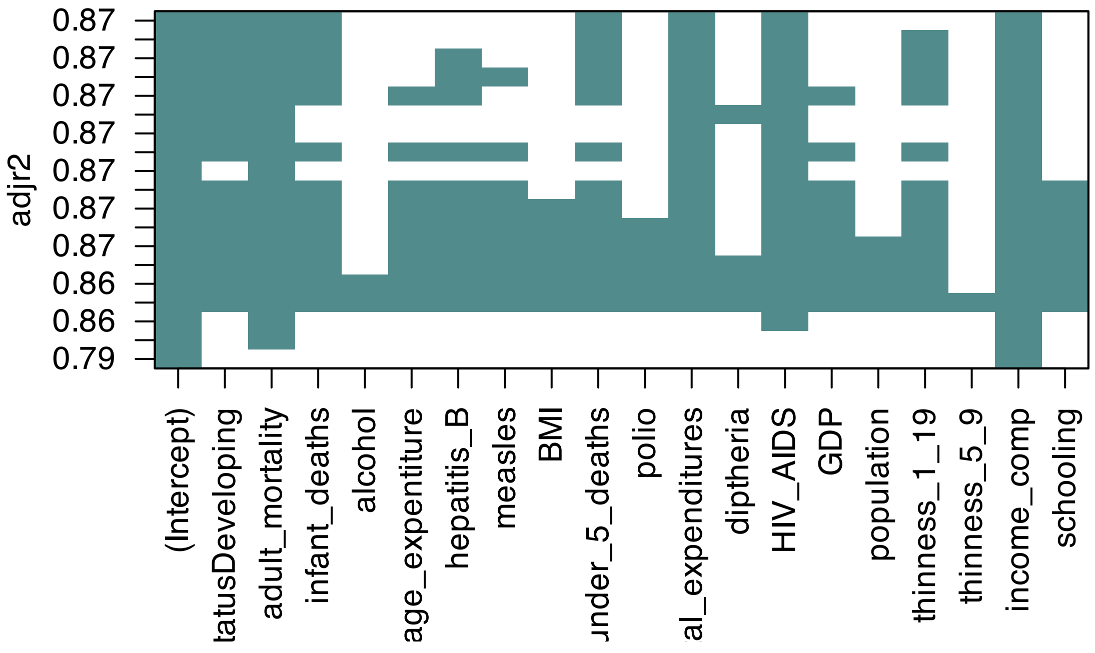


Table 4: Initial models

	Dependent variable:	
	Life Expectancy	
	(1)	(2)
Developing Status	−1.170 (1.035)	-3.395^{***} (1.225)
Adult Mortality	-0.017^{***} (0.004)	-0.035^{***} (0.005)
Infant Deaths	0.083 (0.056)	
Alcohol Consumption	0.006 (0.097)	0.345^{***} (0.109)
Percentage Expenditure	0.0005 (0.0005)	
Hepatitis B	0.012 (0.028)	
Measles	−0.00003 (0.00005)	0.00004 (0.00005)
BMI	−0.008 (0.020)	0.064^{***} (0.023)
Under 5 Deaths	−0.060 (0.038)	−0.002 (0.004)
Polio	−0.009 (0.021)	−0.009 (0.027)
Total Expenditure	0.288^{**} (0.127)	0.329^{**} (0.153)
Diptheria	0.008 (0.034)	0.044^{*} (0.024)
HIV/AIDS	-0.836^{***} (0.247)	-0.924^{***} (0.310)
GDP	−0.0001 (0.0001)	
Population	−0.000 (0.000)	
Thinness 1 to 19	−0.130 (0.227)	
Thinness 5 to 9	0.005 (0.223)	−0.139 (0.118)
Income Composition	35.969^{***} (6.228)	
Schooling	−0.162 (0.274)	
Constant	51.218^{***} (3.314)	71.901^{***} (2.701)
Observations	131	131
R^2	0.883	0.796
Adjusted R^2	0.863	0.777
Residual Std. Error	3.186 (df = 111)	4.062 (df = 119)
F Statistic	44.058^{***} (df = 19; 111)	42.209^{***} (df = 11; 119)
Note:		*p<0.1; **p<0.05; ***p<0.01

Figure 8: Outlier Plot

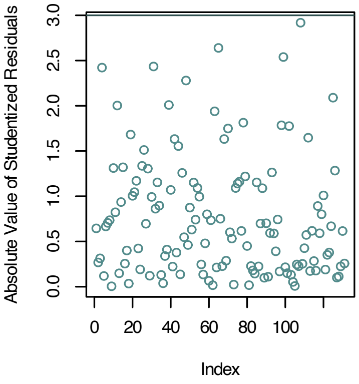


Figure 9: High Leverage Plot

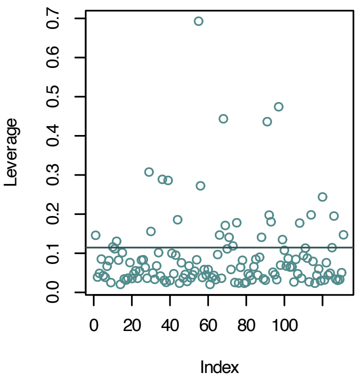


Figure 10: Difference in Fits Plot

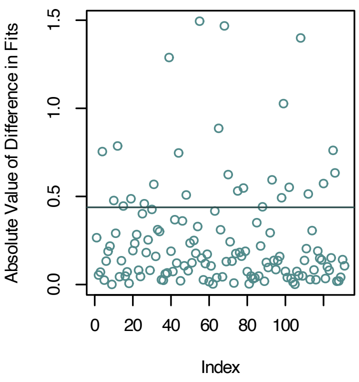


Table 5: Final model

	Dependent variable:
	Life Expectancy
Adult Mortality	-0.016^{***}
	(0.003)
Total Expenditure	0.511^{***}
	(0.094)
HIV/AIDS	-1.142^{***}
	(0.222)
Income Composition	33.193^{***}
	(2.010)
Constant	48.657^{***}
	(1.650)
Observations	124
R^2	0.908
Adjusted R^2	0.905
Residual Std. Error	2.397 (df = 119)
F Statistic	292.725^{***} (df = 4; 119)
Note:	*p<0.1; **p<0.05; ***p<0.01

IV. Final Model

The final model is: Expected life_expectancy = 48.66 - 0.02(adult_mortality) + 0.51(total_expenditure) - 1.14(HIV_AIDS) + 33.19(income_comp)

The r-squared value is 0.9077, indicating that 90.77% of the variation in life expectancy can be explained by this model. The adjusted r-squared value is 0.9046. This is quite high! This suggests that adult mortality, government health expenditure, HIV/AIDS-related deaths among 0 to 4-year-olds, and income composition of resources are useful factors in explaining life expectancy.

The coefficient for adult_mortality is -0.02, meaning that for each one-person increase in adult mortality per 1,000 people, the average expected life expectancy is expected to decrease by 0.02 years. The coefficient for total_expenditure is 0.51, meaning that for each one-percentage increase in government health expenditure, the average expected life expectancy is will increase by 0.51 years. adult mortality per 1,000 people, the average expected life expectancy will decrease by 0.02. The coefficient for HIV_AIDS is -1.14, meaning that for each one-death increase in HIV/AIDS-related deaths among 0 to 4-year olds, the average expected life expectancy will decrease by 1.14 years. The coefficient for income_comp is 33.19, meaning that for each one-point increase in a country's income composition of resources index, the average expected life expectancy will increase by 33.19 years. Since this index ranges only from 0 to 1, the effect of a 0.001 point increase can be obtained from dividing by 1000. This increase is 0.03 years per 0.001 increase in income composition index.

Based on these results, a country aiming to increase its life expectancy should focus on increasing government health expenditure, decreasing HIV/AIDS-related deaths among 0 to 4-year olds, increasing its income composition of resources index, and decreasing adult mortality rates.

V. Conclusion

In this project, a dataset was analyzed with the goal of building a model that accurately explains a country's life expectancy. It was then used to identify key ways in which a country can improve its life expectancy. A model was constructed, indicating that deaths per 1,000 people aged 15 to 60 years, health expenditure as a percentage of total government spending, HIV/AIDS-related deaths among 0 to 4-year-olds per 1,000 births, income composition of resources are key factors in a country's life expectancy. Countries aiming to improve their life expectancy should focus on decreasing the first factor, increasing the second, decreasing the third, and increasing the fourth.

Resources

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download